心理科学进展 2020, Vol. 28, No. 12, 2160–2168
Advances in Psychological Science

https://dx.doi.org/10.3724/SP.J.1042.2020.02160

•研究方法(Research Method)•

认知诊断计算机化自适应测验的选题策略*

唐 倩[1,2] 毛秀珍[1] 何明霜[1] 何 洁[1]

([1]四川师范大学教育科学学院，成都 610066) ([2]德阳市东汽小学，四川 德阳 618000)

摘 要 随着认知诊断计算机化自适应测验(cognitive diagnostic computerized adaptive testing, CD-CAT)理论与实践的发展，兼顾知识状态与能力的双目标 CD-CAT 逐渐受到重视。选题策略是 CAT 的核心，通过梳理传统 CD-CAT 和双目标 CD-CAT 选题策略的研究，并对它们的特点、关系及表现进行介绍和评析。最后，基于认知诊断模型与 CAT 实践发展指出未来应加强一般化认知模型、复杂测验条件认知诊断模型下选题策略的研究；应开发双目标诊断测验的项目和测验特征指标；还应加强非参数选题方法和 CD-CAT 的实践应用研究。

关键词 计算机化自适应测验，认知诊断模型，选题策略，测量精度，非统计约束

分类号 B841

1 引言

项目反应理论(item response theory, IRT)通过分析项目反应数据评估被试连续潜在特质(θ)水平，常用于比较与甄选被试。随着国内、外教育改革的不断深入，教育质量评估要求加强对学生学习过程的形成性评价，并提供详细的教学指导信息以促进教育发展。认知诊断理论(cognitive diagnostic theory, CDT)在教育质量评估实践中应运而生。它能评估被试对特定领域知识的掌握情况、加工技能和认知过程，还能为被试进行补救学习提供个性化的帮助。IRT 和 CDT 作为现代心理与教育测量理论，已广泛应用于分析教育与心理测验数据，并成为计算机化自适应测验(computerized adaptive testing, CAT)的理论基础。

CAT 是一种新型测验模式，实现了测验的量体裁衣。与传统纸笔测验相比，它在获得相似测量精度的条件下既能保证测验效度、测验公平和测验安全，还缩短了测验长度和测验时间，进而提高测验效率(Cheng, 2009)。CAT 自提出以来就倍受研究者和教育实践者的广泛关注。它包括题库、初始项目的选择、选题策略、能力估计方法和测验终止规则几个重要组成部分。若把 CAT 比作一台机器，那么题库便是物质基础，选题策略决定 CAT 的运转方式，能力估计方法是推动力，终止规则就是停止键。其中，选题策略决定了单个项目的适切性，也决定着整个测验的效率和测验公平，还是影响测验成本和测验安全的重要因素。于是，选题策略成为 CAT 研究的核心内容之一，影响着 CAT 未来发展的方向。

起初，CAT 仅评估被试宏观的单维(多维)连续潜在特质水平(θ)或者微观离散的知识状态(knowledge states, KS) (α)。虽然连续潜在特质与离散 KS 代表被试不同侧面的特征，但它们并不相互排斥，而是密切联系的统一体。因而，CAT 中如何同时评估被试的 θ 和 α 成为一个有价值的研究问题，推动了兼顾 KS 和能力的双目标 CD-CAT (dual-objective CD-CAT)的研究。于是，本文首先系统梳理了传统 CD-CAT (以估计被试 KS 为目的的 CD-CAT)和双目标 CD-CAT 选题策略的研究进展。然后，通过对各类选题策略的特点、关系及表现进行介绍和评价，以期把握其发展脉络和趋势。最后，基于认知诊断模型与 CAT 实践指出未来研究的几个方向：应加强一般化认知模型、复杂测验条件认知诊断模型下选题策略的研究；应开发双目标诊断测验的项目和测验特征指标；还

收稿日期：2020-02-06
* 国家自然科学基金青年项目(31400897)。
通信作者：毛秀珍, E-mail: maomao_wanli@163.com

应加强非参数选题方法和 CD-CAT 的实践研究。

下文用 K、L 和 T 分别表示测验考察的属性个数、测验长度和已施测项目数，$\{i_1,i_2,\cdots,i_T\}$ 与 $X_T=(x_1,x_2,\cdots,x_T)$ 代表已施测项目及反应，R 与 $R_T=R/\{i_1,i_2,\cdots,i_T\}$ 表示题库和剩余题库，α 和 $\hat{\alpha}$ 分别表示真实的和当前估计的 KS。另外，所有 KS 的集合为 $\Omega=\{\alpha_1,\alpha_2,\cdots,\alpha_{2^K}\}$，所有 KS 对的集合为 $\Omega_0=\{(\alpha_u,\alpha_v)\,|\,\alpha_u,\alpha_v\in\Omega\}$。$\Omega_0$ 的子集 $\Omega_1=\{(\alpha_u,\alpha_v)\,|\,\alpha_{uk}\neq\alpha_{vk},\alpha_{um}=\alpha_{vm},k\in\{1,2,\cdots,K\},m\neq k\}$ 表示仅在某一个属性上具备不同掌握状态的 KS 对的集合。集合 $\Omega_{1k}=\{(\alpha_u,\alpha_v)\,|\,\alpha_{uk}=1,\alpha_{vk}=0,\forall m\neq k,\alpha_{um}=\alpha_{vm}\}$ 是 Ω_1 的子集，仅与属性 k 相关。

2 传统 CD-CAT 项目选择策略

测量精度是 CAT 的首要目标，但 CAT 又不能一味追求高测量精度而不顾一些非统计约束条件，如内容约束和项目曝光控制。因而，CAT 发展至今，其选题策略的研究几乎都围绕这些方面展开。

2.1 提高测量精度的选题方法

项目特征与被试 KS 是 CAT 选择项目的依据。近二十年来，研究者不断突破传统 CAT 的研究思路，极大地丰富了 CD-CAT 选题策略的研究。总体上，传统 CD-CAT 选题策略沿着：项目反应分布、KS 后验分布和结合项目与被试特征视角建构提高测量精度的选题指标。

2.1.1 基于项目反应分布的信息量选题指标

Kullback-Leibler (KL)是最基础的选题指标。项目 j 的 KL 信息量等于反应分布 $P(x_j|\hat{\alpha})$ 和 $P(x_j|\alpha_c)(c=1,2,\cdots 2^K)$ 的 KL 信息量之和，即 $KL_j=\sum_{c=1}^{2^K}KL_j(\hat{\alpha}\|\alpha_c)$ (Xu et al., 2003)。利用 KS 的后验分布 $P(\alpha_c|X_t)$ 对 KL 信息加权就得到后验加权 KL 信息(posterior-weighted KL, PWKL)，$PWKL_j=\sum_{c=1}^{2^K}KL_j(\hat{\alpha}\|\alpha_c)\cdot P(\alpha_c|X_t)$。如果进一步利用海明距离 $h(\hat{\alpha},\alpha_c)$ 反映 $\hat{\alpha}$ 与 α_c 的相似性，便得到 $HKL_j=\sum_{c=1}^{2^K}KL_j(\hat{\alpha}\|\alpha_c)\cdot P(\alpha_c|X_t)\cdot(h(\hat{\alpha},\alpha_c))^{-1}$，称为混合 KL 信息(the hybrid KL, HKL)。最后，若 PWKL 中 $\hat{\alpha}$ 取遍所有可能情况，就是修订的 PWKL 方法，记为 $MPWKL_j=\sum_{d=1}^{2^K}P(\hat{\alpha}_d|X_t)\cdot PWKL_j(\hat{\alpha}_d)$。

KL、PWKL、HKL 和 MPWKL 均选择信息量最大的项目。它们是几种最基础和常用的选题策略。研究表明，PWKL 和 HKL 表现类似，均优于 KL 方法(Cheng, 2009; Wang, 2013)。与 PWKL 相比，MPWKL 计算更复杂，短测验测量精度更高。当测验长度达 20 及以上时二者无明显差异(Kaplan et al., 2015)。

2.1.2 基于 KS 后验分布的信息量选题指标

香农熵(shannon entropy, SHE)和互信息(mutual information, MI)是基于 KS 后验分布的选题方法。前者选择使预测 $\hat{\alpha}$ 后验分布的期望香农熵最小的项目，后者选择使 $P(\alpha|X_T)$ 与 $P(\alpha|X_T,x_{T+1})$ 的预测 KL 信息最大的项目(Zheng & Chang, 2016)。换言之，

$$SHE=\arg\min_{j\in R_T}\left\{\sum_{x=0}^{1}H(P(\alpha_c\,|\,X_T,x_j=x))\cdot P(x_j=x\,|\,X_T)\right\}, \quad (1)$$

$$MI=\arg\max_{j\in R_T}\left\{\sum_{x=0}^{1}P(x_j=x|X_T)\cdot\sum_{c=1}^{2^K}KL(P(\alpha_c\,|\,X_T,x_j=x)\,\|\,P(\alpha_c\,|\,X_T))\right\}. \quad (2)$$

Wang (2013)指出在大多数情况下 MI 方法比 PWKL、SHE 和 KL 的测量精度更高。由于 SHE 和 MI 方法涉及预测反应分布，计算比较复杂，他将 MI 展开并进行简化得到简化 MI 方法。该方法计算更简单、所需时间更短且不明显降低测量精度。

2.1.3 结合项目与被试特征的选题指标

认知诊断模型中项目特征包括 q 向量、认知诊断区分度(cognitive discrimination index, CDI) (Henson & Douglas, 2005)、属性区分度(attribute discrimination index, ADI) (Henson et al., 2008)和广义决定性输入，噪音“与”门模型(the generalized deterministic inputs, noisy “and” gate model, G-DINA)区分度指标 ξ^2 (de la Torre & Chiu, 2016)，被试特征主要指 KS 的后验概率 $P(\alpha_c\,|\,X_T)$ $(c=1,2,\cdots,2^K)$ 和属性掌握概率 $P(\alpha_{ik}=1|X_T)(k=1,2,\cdots,K)$。

一方面，基于 $P(\alpha_c\,|\,X_T)$ 与项目 q 向量建构了二分法(halving algorithm, HA) (汪文义等, 2014)，结合 $P(\alpha_c\,|\,X_T)$ 与 ξ^2、CDI 和 ADI 提出 G-DINA

模型区分度指标(the G-DINA model discrimination index, GDI)选题方法(Kaplan et al., 2015)、后验加权 CDI (posterior-weighted CDI, PWCDI)和后验加权 ADI (posterior-weighted ADI, PWADI)方法(Zheng & Chang, 2016)，分别见式(3)~(6)。

$$HA = \arg\min_{j \in R_T} \left\{ \left| \sum_{c:\alpha_c' q_j \geqslant q_j' q_j} P(\alpha_c \mid X_T) - 0.5 \right| \right\} \quad (3)$$

$$GDI = \arg\max_{j \in R_T} \left\{ \sum_{c=1}^{2^K} P(\alpha_c \mid X_T)[P(x_{ij}=1 \mid \alpha_c) - \sum_{c=1}^{2^K} P(\alpha_c \mid X_T) P(x_{ij}=1 \mid \alpha_c)]^2 \right\}, \quad (4)$$

$$PWCDI = \arg\max_{j \in R_T} \{ (\sum_{\Omega_0} h(\alpha_u, \alpha_v)^{-1})^{-1} \sum_{\Omega_0} h(\alpha_u, \alpha_v)^{-1} P(\alpha_u | X_T) \cdot P(\alpha_v | X_T) \cdot KL_j(\alpha_u \| \alpha_v) \}, \quad (5)$$

$$PWADI = \arg\max_{j \in R_T} \{ (K \cdot 2^K)^{-1} \cdot \sum_{\Omega_1} P(\alpha_u \mid X_T) \cdot P(\alpha_v \mid X_T) \cdot KL_j(\alpha_u \| \alpha_v) \}. \quad (6)$$

值得注意的是，MPWKL 是所有 KS 对 (α_u, α_v) 的 KL 信息量与 α_u 、α_v 后验概率乘积之和，实质上与汪文义等(2014)的 KLED 方法等价。PWCDI 本质上是对 MPWKL 各项取加权平均，PWADI 则是 MPWKL 中特定项的平均值。PWCDI 最复杂，PWADI 最简单，研究发现它们的模式判准率无明显差异(Zheng & Chang, 2016)。

另一方面，项目与被试特征指标还常作为权重对信息量加权构建选题指标。例如，郭磊等(2016)运用 CDI、ADI 对 PWKL 信息量加权得到 CDIPWKL 和 ADIPWKL 方法，能提高 PWKL 的测量精度。又如，罗照盛等人(2015)利用边际属性掌握概率之差 $\sum_{k=1}^{K} | P(\alpha_{ik}=1 | X_T, x_j) - P(\alpha_{ik}=1 | X_T) |$ 对 PWKL 和 HKL 加权得到 PPWKL 和 PHKL 方法，也能提高 PWKL 和 HKL 在测量精度和项目曝光均匀的综合表现。此外，研究者还分析特定模型的项目特征指标，如 DINA 模型的项目鉴别力指数：高分组的通过率(1 与失误参数 s_j 之差)减去低分组通过率(猜测参数 g_j)，记为 $w_j = 1-(s_j + g_j)$ (Rupp et al., 2010)，并作为 DINA 模型下项目信息量的加权指标(郭磊等，2016)。由此可见，除了一般项目特征指标外，研究特定模型下的项目特征也具有重要意义。

2.2 属性平衡的选题策略

认知属性是诊断测验的显著特点，也是最小的内容单元。平衡属性考察次数是保证测验效度的关键。

2.2.1 最大优先指标(maximum priority index, MPI)方法

最大优先指标 $MPI = \prod_{k=1}^{K} [(u_k - b_k)/u_k]^{q_{jk}}$ 结合了属性 k 的目标最大测量次数 u_k 、当前已考察次数 b_k 和 Q 矩阵元素 q_{jk} (Cheng, 2010)。运用 MPI 对项目信息量加权选题可以提高测量精度。事实上，$(u_k - b_k)/u_k$ 的值总小于等于 1。于是，MPI 的值随项目考察的属性增多而减小，并倾向于选择考察属性较少的项目，导致项目曝光不均匀。鉴于此，余丹等人(2011)、刘舒畅等人(2018)、孙小坚等人(2019)对 MPI 进行修订，分别提出 $MPI_1 = \prod_{k=1}^{K} [(u_k - b_k)/u_k + 1]^{q_{jk}}$ 、$MPI_2 = \sum_{k=1}^{K} [(u_k - b_k)/u_k]^{q_{jk}}$ 和 $MPI_3 = \left[\prod_{k=1}^{K} (u_k - b_k)^{q_{jk}} \right] \Big/ C, (C > 0)$。此外，刘舒畅等人(2018)利用当前(目标)标准误 SE_{bk} (SE_{Bk}) 建立了 $MPI_4 = \sum_{k=1}^{K} [(SE_{bk} - SE_{BK}) / SE_{BK}]^{q_{jk}}$ ；孙小坚等人(2019)还将 Kuo 等人(2016)针对测验建构提出的平衡属性模式的权重指标 RTA 用于满足属性的最少测量次数。其中，$RTA_j = (1 + I((T/K) < 3) \sum_{t=1}^{T} I(q_j = q_t))^{-1}$, $I(\cdot)$为指示函数，q_j 和 q_t 分别是未作答和已作答题目的 q 向量。

刘舒畅等人(2018)和孙小坚等人(2019)系统考察了各个优先指标与 CDI、KL、PWKL、MPWKL 和 MI 乘积的选题表现。结果一致表明，改进的优先指标比 MPI 的测量精度更高。大部分测验条件下，MPI_4 优于 MPI_2 , MPI_2 优于 MPI_1 ； MPI_3 与 MPI_2 与不同选题策略结合选题各有优势。一般而言，MPI_3 较 MPI_2 更能平衡项目曝光，测量精度稍低。

2.2.2 基于加权离差思想构建的选题方法

Lin 和 Chang (2018)借鉴加权离差模型

(Swanson & Stocking, 1993), 建立了属性偏差指标 $WD_j = \sum_{k=1}^{K} w_k(l_k - b_k - q_{jk}) + \sum_{k=1}^{K} w_k(u_k - b_k - q_{jk})$ 和标准化加权属性偏差指标 $SWD_j = \frac{Max(WD_j) - WD_j}{Max(WD_j) - Min(WD_j)}$。$w_k$ 为属性 k 的权重, WD_j 只计算每个属性与其上、下界的正离差的加权和。类似地, KL 可标准化为 $SKL_j = \frac{KL_j(\hat{\alpha}_i) - Min(KL_j(\hat{\alpha}_i))}{Max(KL_j(\hat{\alpha}_i)) - Min(KL_j(\hat{\alpha}_i))}$。他们比较了 $(-WD_j) \cdot KL_j$ (记为 WDKL)和 $SWD_j \cdot SKL_j$ (记为 SWDKL)的结果, 指出 SWDKL 虽然在平衡属性测量次数和模式判准率方面比 WDKL 表现更好, 但它们的项目曝光不均匀。

2.3　曝光控制的选题方法

针对项目曝光不均匀性问题, 研究者考察了传统 CAT 中限制阈值方法(restrictive threshold, RT)、限制进度方法(restrictive progressive, RPG)、分层方法和最大优先指标的表现(Wang et al., 2011; 毛秀珍, 辛涛, 2013)。Lin 和 Chang (2018)还对 RPG 适当变形并结合 SWDK 和优先指标, 提出约束渐进的 SWDKL 方法(the constrained progressive SWDKL, CP_SWDKL):

$$CP_SWDKL_j(\hat{\alpha}_i) = \frac{er_{\max}}{er_j}\left[\left(1 - \frac{T}{L}\right)R_j + \frac{T}{L} \times R_{jl}\right] \quad (7)$$

$er_{\max}$ 与 er_j 表示要求的最大曝光率和项目 j 的曝光率, s 是调整 R_{jl} 区间长短的量, 值越小, 区间越大, 选题越随机。令 $a = \min\{SWDKL_j, j \in R_T\}$、$b = \max\{SWDKL_j, j \in R_T\}$, 随机数 $R_j \in U(a,b)$, $R_{jl} \in U(SWDKL_j - (SWDKL_j - a)/s, SWDKL_j + (b - SWDKL_j)/s)$。研究表明, CP_SWDKL 能显著提高 SWDKL 和 KL 的项目曝光均匀性, 但也在一定程度上降低测量精度。总体上讲, RT 和 RPG 方法能较好地控制项目曝光率并提高题库利用率。

2.4　CD-CAT 选题策略简评

除了依据测量目的外, 还可以从选题方法的建构思路对传统 CD-CAT 的选题策略分类(见表 1)。对选题策略的研究, 有以下几点思考。第一: 除依据 KS 后验分布定义香农熵和互信息外, 还可以运用其它特征变量, 如预测反应分布建立香农熵和互信息选题方法。鉴于 SHE 和 MI 等方法计算复杂, 研究简化基于 KS 后验分布的选题方法、挖掘它们的关系都具有重要意义。第二, 属性偏差指标是各个属性测量次数离差的加权和, 优先指标是各个属性测量次数离差与目标占比的等权重加权和, 二者实质上具有相同的建构思路。因此, 基于属性其它特征, 如测量信息量离差建立加权指标也是建构选题方法的一种重要思路。第三, 加权选题方法集中在对反应分布信息量指标的加权, 研究适合其它基础选题指标的加权方法也是未来有价值的研究问题。最后, 结合多种思路的方法主要解决项目曝光不均匀问题, 但大部分研究局限于传统 CAT 的思想, 缺乏系统对比。因此, 基于认知诊断测验的特点发展结合多种思路的选题方法是今后研究的重点。

传统 CAT (CD-CAT)在测验结束时只报告 $\hat{\theta}$ ($\hat{\alpha}$)。兼顾 KS 和能力的双目标 CD-CAT 能同时评估 $\hat{\alpha}$ 和 $\hat{\theta}$, 引领 CD-CAT 的发展方向, 具有重要的实践价值。

表 1　传统 CD-CAT 选题策略汇总表

分类标准	特点	具体方法	适用情景
基础选题指标	反应分布信息量指标	KL、PWKL、HKL、MPWKL	提高测量精度
	KS 后验分布信息量指标	SHE、MI	
	基于项目、被试特征选题	HA、GDI、PWCDI、PWADI	
加权选题方法	基于区分度、KS 后验概率加权	CDIPWKL、ADIPWKL、PPWKL、PHKL	
	优先指标加权: MPI 及变式 $MPI_i (i = 1,2,3,4)$	对信息量(KL、PWKL、MPWKL、MI)加权; $MPI_1 \cdot CDI$ 、 $MPI_2 \cdot CDI$	平衡属性测量次数
	属性偏差指标加权	WDKL、SWDKL	
结合多种思路	运用多个步骤或方法	RT、RPG、分层方法、优先指标法、P-SWDKL	平衡项目曝光率

3 双目标 CD-CAT 的项目选择策略

如何表征项目关于能力与 KS 的信息是双目标 CD-CAT 选题策略的核心，也是区别于传统 CD-CAT 的重要特征。根据 KS 与能力信息量的结合方式，可将双目标 CD-CAT 的选题方法分为三类：两阶段选题法、信息量加权平均方法和约束加权信息量方法。

3.1 两阶段选题方法

两阶段选题方法包括两步法和影子测验方法。首先，两步法在传统 CAT (或 CD-CAT)测验结束时利用所有项目的反应估计 α (或 θ)，是实现双目标 CD-CAT 最直接的方法。其次，CAT 中影子测验方法在选题之前依据一定标准构造影子测验，然后在影子测验中选择最优项目，通过两步选题为实现双目标 CD-CAT 提供了可能。例如，McGlohen 和 Chang (2008)、杜宣宣(2010)分别以 $\hat{\theta}$ 和 $\hat{\alpha}$ 构造影子测验，然后分别选择使 $\hat{\alpha}$ 和 $\hat{\theta}$ 信息量最大的项目。

两步法简单易行，但仅依据 α (或 θ)的信息选题，不能同时保证 $\hat{\alpha}$ 和 $\hat{\theta}$ 的测量精度(McGlohen & Chang, 2008)。与两步法相比，影子测验方法能有效提高 α 和 θ 的估计精度，还提高了项目曝光均匀性。但影子测验方法也是将 α 和 θ 的信息独立地应用于选题，往往只能优先保证 α 或 θ 的估计精度。因此，结合 α 和 θ 的信息建立项目选择指标成为探索双目标 CD-CAT 研究的新方向。

3.2 信息量加权平均方法

3.2.1 双信息选题方法(dual information, DI)

Cheng(2007)首次依据 $PWKL_j(\hat{\alpha})$ 和 $KL_j(\hat{\theta})$ 提出项目 j 的 DI 信息量指标 $DI_j = wPWKL_j(\hat{\alpha}^T) + (1-w)KL_j(\hat{\theta}^T)$。DI 方法结合了 KS 和能力估计值的信息，选择使 DI 值最大的项目。但是 $PWKL_j(\hat{\alpha})$ 和 $KL_j(\hat{\theta})$ 的取值相差较大，后者对 DI 的影响很小(Wang et al., 2014)。于是，将它们转换到相同量表再加权平均无疑是一种更合理的方法。

3.2.2 信息量统一量纲加权平均方法

百分等级、标准化转换与对数值转换是统计学上常用的统一量纲方法。鉴于此，Wang 等人(2014)提出先对 $PWKL(\hat{\alpha})$ 和 $PWKL(\hat{\theta})$ 进行百分等级($pe[\cdot]$)或标准分数($s[\cdot]$)转换后再加权平均，得到百分等级合成法(aggregate ranked information method, ARI)和标准差合成法(aggregate standardized information method, ASI)，即：

$$ARI_j = \arg\max\{w \cdot pe[PWKL_j(\hat{\alpha}^T)] + (1-w) \cdot pe[PWKL_j(\hat{\theta}^T)], j \in R_T\} \quad (8)$$

$$ASI_j = \arg\max\{w \cdot s[PWKL_j(\hat{\alpha}^T)] + (1-w) \cdot s[KL_j(\hat{\theta}^T)], j \in R_T\} \quad (9)$$

不同于百分等级和标准化思想，Dai 等人(2016)提出先对 $\hat{\theta}$ 的 Fisher 信息量 $I_j(\hat{\theta})$ 和 $SHE_j(\hat{\alpha})$ 进行对数值转换，然后加权求和得“带有信息量的有序度” (dapperness with information, DWI)选题指标 $w\log(I_j(\hat{\theta})) - (1-w)\log(SHE_j(\hat{\alpha}))$。他们通过模拟研究表发现 ω 取 0.5 时表现最佳。于是，DWI 方法便简化为选择使 $I_j(\hat{\theta}) / SHE_j(\hat{\alpha})$ 最大的项目。研究表明，DWI 与影子测验方法相比，模式判准率相似，能力估计更准确。

由此可见，ARI、ASI 与 DWI 都基于 DI 方法对 $\hat{\alpha}$ 和 $\hat{\theta}$ 的信息量进行量纲统一转换而来。ARI 对连续信息量排序获得对应的百分等级，在一定程度上导致信息丢失，且容易受题库大小的影响；DWI 能避免题库大小和极端值对选题的影响(Zheng et al., 2018)。总体上讲，与 DI 方法相比，在 KS 和能力估计方面，ARI 的表现更差，ASI 方法更优，DWI 方法的能力估计精度更高。此外，ASI 和 ARI 方法不局限于 PWKL 信息量。例如，Kang 等人(2017)在 ASI 和 ARI 方法中运用 MPWKL 信息量后分别得到 MASI 和 MARI 方法。

3.2.3 Jensen-Shannon (JS)距离选题方法

根据 Lin (1991)中 JS 距离的定义，Kang 等人(2017)首先定义加权分布 $g = wf_{\hat{\alpha}} + (1-w)f_{\hat{\theta}}$，然后对 $KL(f_{\hat{\alpha}} \| g)$ 与 $KL(f_{\hat{\theta}} \| g)$ 求加权平均，进而定义项目 j 的 JS 距离如下：

$$JS_j(f_{\hat{\alpha}} \| f_{\hat{\theta}}) = wKL_j(f_{\hat{\alpha}} \| g) + (1-w)KL_j(f_{\hat{\theta}} \| g). \quad (10)$$

JS 距离满足非负、对称和三角不等式性质。不同于 KL、MI 和 SHE，还可以定义有限个概率分布的 JS 距离，并且允许根据各个概率分布的重要性加权。研究表明，JS 方法通过选择使 JS 距离最大的项目，其模式判准率明显高于 ARI、ASI、MARI 和 MASI 方法，而且 JS 方法的项目曝光更均衡，计算时间更短。此外，Kang 等人(2017)还基于香农熵定义了 JS 距离，并研究了 JS 距离与互信息、Fisher 信息量的关系。

3.3　约束加权信息量方法

Wang 等人(2012)和 Zheng 等人(2018)都指出双目标 CD-CAT 中可将认知诊断判准率视作内容约束，分别提出加权信息量方法和信息量乘积选题方法(information product approach, IPA)。Wang 等人(2012)考虑到优先指标 MPI 中 q_{ik} 作为指数导致求和或求积的项数等于项目考察的属性个数从而带来不可比问题。于是，他们改变 q_{jk} 的位置提出 Q 矩阵控制指标：$P_1=\prod_{k=1}^{K}[(u_k-b_k-q_{jk})/u_k]\cdot[((L-l_k)-(t-b_k-q_{jk}))/(L-l_k)]$、KL 信息控制指标 $P_2=\sum_{k=1}^{K}(u_k-b_k)\left[\sum_{\Omega_{1k}}KL(\alpha_u,\alpha_v)/2^{K-1}\right]$ 和 DINA 模型 Q 区分控制指标 $P_3=(1-s_j)(1-g_j)\cdot P_1$，然后分别对 $\hat{\theta}$ 的 Fisher 信息量加权选题。其中，l_k 是属性 k 的目标最小测量次数。与 P_1 和 P_3 相比，P_2 加权选题对 KS 和 θ 的估计精度都最高(Wang et al., 2012)。注意到，除 P_3 仅适用于 DINA 模型外，其它优先指标可用于任何诊断模型。由此，针对特定模型提出切合模型特点的指标同样具有重要意义。

IPA 方法将认知诊断项目信息量视作极大优先指标并与能力信息量相乘而得。Zheng 等人(2018) 考察了 $PWKL_j(\hat{\alpha})\cdot KL_j(\hat{\theta})$ 与 $PWADI_j\cdot KL_j(\hat{\theta})$ 的选题表现，指出 IPA 方法对 α 和 θ 的估计比 ASI 和 ARI 更准确，同时 IPA 没有权重要求，不受题库和极端值的影响。特别地，DWI 方法中 $1/SHE(\hat{\alpha})$ 可视为极大优先指标，从而 DWI 方法也是一种 IPA 方法。另外，若对 IPA 方法取对数就转换为信息量对数之和，即 $\log IPA=\log PWKL(\hat{\alpha})+\log KL(\hat{\theta})$，这又成为 $\log PWKL(\hat{\alpha})$ 和 $\log KL(\hat{\theta})$ 的加权平均。因此，IPA 在一定程度上具备双目标 CD-CAT 项目选择方法的一般性框架。

3.4　双目标 CD-CAT 选题策略简评

两阶段方法、信息量加权平均方法和约束加权信息量方法是三类双目标 CD-CAT 选题策略，见表 2。首先，两阶段选题方法将 α 和 θ 的信息量独立地应用于选题。于是，将测验按比例分成多个阶段或者结合两者信息建构影子测验都可能提高两阶段方法选题表现。其次，信息量加权平均方法创新性地将 α 和 θ 的信息量统一为一个选题指标，但二者取值相差较大。于是，研究者一方面运用百分等级、标准分数、对数转换改进信息量加权平均方法，另一方面通过对 α 和 θ 的反应分布加权来建立 JS 距离选题方法。信息量加权平均方法主要运用了 CD-CAT 中常用的 PWKL、KL 和 SHE 选题指标，并且表现较好的 DWI、ARI 和 JS 方法在大部分测验条件下对 KS 的判准率在 0.9 左右，RMSE 在 0.4 左右(Wang et al., 2014; Dai et al., 2016; Kang et al., 2017)，测量精度不够高。因此，今后还应考察多种信息量指标、开发双目标 CD-CAT 项目特征指标等方式研究双目标 CD-CAT 选题策略，提高测量精度。

注意到，权重是信息量加权平均方法的重要部分。通过比较 0 到 1 之间多个权重，Cheng(2007)指出除极端权重值外，不同权重对 DI 方法的影响很小，Dai 等人(2016)发现权重为 0.5 时，DWI 方法表现最优。Wang 等人(2014)则系统对比了三类权重指标。第一，理论的权重，即选择第 t 个项目时权重为 $w=t/(L+1)$；第二，实证的权重，即基于累积信息量 $Inf_{\theta}^{(T)}$ 和 $Inf_{\alpha}^{(T)}$ 与目标信息量 Inf_{θ} 和 Inf_{α} 的差距占目标信息量比重，如 $w_{\theta}=w_1/(w_1+w_2)$ (其中 $w_1=(Inf_{\theta}-Inf_{\theta}^{T})/Inf_{\theta}$，$w_2=(Inf_{\alpha}-Inf_{\alpha}^{T})/Inf_{\alpha}$)；第三，通过属性的权重向量 $(\tau_1,\tau_2,\cdots\tau_K)$ 与属性水平信息量向量的数量积构造属性层面的权重。他们指出，ASI 和 ARI 方法中运用理论或实证权重都优于等权重。理论权重适用于高质量题库，实证权重适用于信息量较少的题库，属性层面的权重则适用于属性具有不同权重的情况。

表 2　双目标 CD-CAT 选题策略汇总表

两阶段选题	信息量加权平均	约束加权信息量
两步法	直接加权平均：DI	Q 矩阵、KL 信息控制指标加权：$P_1\cdot FI(\hat{\theta})$、$P_2\cdot FI(\hat{\theta})$、$P_3\cdot FI(\hat{\theta})$
影子测验方法	统一量纲：ASI、ARI、MASI、MARI、DWI 分布反应加权：JS	信息量乘积：IPA

再次，约束加权信息量方法可以视作 CD-CAT 加权信息量方法的扩展。特别地, IPA 方法经对数转换可视为信息量加权平均方法，而加权平均 DWI 方法又可视为 IPA 方法。因此，IPA 方法具有双目标 CD-CAT 选题策略的一般性框架。最后，双目标 CD-CAT 选题策略集中于提高测量精度的研究，而项目曝光均匀性和内容约束相比于传统 CD-CAT 都具有新的特点和挑战。因此，今后可以借鉴传统 CD-CAT 中选题策略的思路和方法，结合项目特征、KS 和能力的信息建构双目标 CD-CAT 选题策略，结合多种方法研究具有非统计约束的选题方法。

4 研究展望

CD-CAT 自提出以来，因其对知识结构的诊断功能和 CAT 的高效测验模式，得到研究者的广泛关注和深入研究。特别地，针对 CD-CAT 选题的测量精度、项目曝光和内容约束问题，研究者不仅将传统 CAT 的选题策略推广到 CD-CAT，还基于认知诊断测验的特征发展了独特的选题方法。不仅如此，随着研究深入和实践需要，兼顾能力和 KS 的双目标 CD-CAT 也得到广泛关注，并有大量研究。传统 CD-CAT 和双目标 CD-CAT 结合了 IRT、CDT 和 CAT 的理论与技术。它们的发展与测量理论的研究与实践、计算机技术的发展密切相关。

首先，近 20 年来认知诊断模型得到了极大的丰富和发展，呈现出从单一测验条件到复杂测验条件，从低阶到高阶，从特殊到一般的发展特点。一方面针对二级评分项目提出了一般化 G-DINA 模型。它在一定约束条件下可得到 DINA、DINO、NIDA、NIDO、RUM 和 ACDM。然而，目前 CD-CAT 研究还以约束化认知诊断模型为基础，并以 DINA 模型和 RUM 模型为主。因此，基于一般诊断模型研究 CD-CAT 具有重要意义。这不仅能统一不同模型下项目选择和能力估计算法的编码过程，还有利于比较它们在不同模型下的表现。

另一方面，认知诊断模型还围绕 G-DINA 和约束化诊断模型扩展了复杂测验条件模型，如多级评分、属性多级和高阶模型。当前 CD-CAT 以二级评分项目为主，并有少量多级评分项目的研究。于是，探索多策略、属性多级评分和项目多级评分甚至更复杂测验条件下 CD-CAT 选题策略同样是今后研究的重要方向。

其次，CD-CAT 中结合项目和被试特征是改进选题策略的重要思路。于是，针对双目标 CD-CAT，如何构建表征能力和认知特征的项目与测验特征指标，如区分度指标；如何基于双目标认知诊断测验项目特征构建选题策略都是具有意义的研究问题。此外，目前的选题方法各有优势与不足，有必要探讨它们的最佳组合模式，加强非统计约束选题策略的研究。

最后，CD-CAT 在国内实践还处于起步阶段，仅 2009~2011 年教育部组织了数学和英语的大规模 CD-CAT 测试(Liu et al., 2013)。因此，今后有必要研究非参数项目选择方法，既可用于小规模课堂诊断实践，还能为大规模实践应用收集数据做准备。

参考文献

杜宣宣. (2010). *具有认知诊断功能的计算机化自适应测验的选题策略研究* (硕士学位论文). 江西师范大学，南昌.

郭磊，郑蝉金，边玉芳，宋乃庆，夏凌翔. (2016). 认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标. *心理学报, 48*(7), 903–914.

刘舒畅，涂冬波，蔡艳，赵洋. (2018). 四种新的基于属性平衡的 CD-CAT 选题策略开发研究. *心理科学, 41*(4), 976–981.

罗照盛，喻晓锋，高椿雷，李喻骏，彭亚风，王睿，王钰彤. (2015). 基于属性掌握概率的认知诊断计算机化自适应测验选题策略. *心理学报, 47*(5), 679–688.

毛秀珍，辛涛. (2013). 认知诊断 CAT 中项目曝光控制方法的比较. *心理学报, 45*(6), 694–703.

孙小坚，王钰彤，张世夷，辛涛. (2019). 认知诊断计算机自适应测验中平衡属性收敛的新方法. *心理科学, 42*(5), 1236–1244.

汪文义，丁树良，宋丽红. (2014). 兼顾测验效率和题库使用率的 CD-CAT 选题策略. *心理科学, 37*(1), 212–216.

余丹，潘奕娆，丁树良，杨庆红. (2011). 计算机化自适应诊断测验新的选题策略. *江西师范大学学报(自然科学版), 35*(5), 548–550.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*(6), 902–913.

Cheng, Y. (2007). *The dual information method for item*

selection in cognitive diagnostic computerized adaptive testing (Master's thesis). University of Illinois at Urbana-Champaign.

Dai, B. Y., Zhang, M. Q., & Li, G. M. (2016). Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: Based on the RRUM. *Applied Psychological Measurement, 40*(8), 625–640.

de la Torre, J., & Chiu, C. Y. (2016). A General method of empirical Q-Matrix validation. *Psychometrika, 81*(2), 253–273.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262–277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

Kang, H.-A., Zhang, S., & Chang, H.-H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 54*(2), 165–183.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188.

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement, 40*(5), 315–330.

Lin. C.-J., & Chang, H.-H. (2018). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement, 79*(2), 335–357.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of classification, 30*(2), 152–172.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808–821.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* New York: Guiford Press.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035.

Wang, C., Chang, H. -H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44*(1), 95-109.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48*(3), 255–273.

Wang, C., Zheng, C., & Chang, H.-H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement, 51*(4), 358–380.

Xu, X. L., Chang, H. H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*(8), 608–624.

Zheng, C., He, G., & Gao, C. (2018). The information product methods: A unified approach to dual-purpose computerized adaptive testing. *Applied Psychological Measurement, 42*(4), 321–324.

Item selection methods for cognitive diagnostic computerized adaptive testing

TANG Qian[1,2], MAO Xiuzhen[1], HE Mingshuang[1], HE Jie[1]

([1] *Institute of Educational Science, Sichuan Normal University, Chengdu* 610066, *China*)

([2] *Dongqi Primary School, Deyang,* 618000, *China*)

Abstract: Dual-objective cognitive diagnostic computerized adaptive testing (CD-CAT), which considers knowledge status and ability simultaneously, has become more and more popular with the theoretical and practical development of CD-CAT. Item selection methods play a key role in CD-CAT. This paper systematically reviews existing item selection methods on traditional and dual-objective CD-CAT, and summarizes the types, characteristics, relations, and performance of these methods. Furthermore, several future research directions were illustrated. First, it is necessary to study item selection strategy with general cognitive models and under complex test conditions. Second, it is important to develop indexes representing items and test characteristic of dual-objective diagnostic testing. Finally, it is meaningful to conduct research on non-parametric item selection methods and practical applications of CD-CAT.

Key words: computerized adaptive testing, cognitive diagnostic model, item selective strategy, measurement accuracy, non-statistical constraints